DOCUMENT RESUME

ED 298 863 HE 021 854

AUTHOR Cook, Stuart S.
TITLE Improving the Quality of Student Ratings of Instruction: A Look at Two Strategies. AIR 1988 Annual Forum Paper.
PUB DATE May 88
NOTE 26p.; Paper presented at the Annual Forum of the Association for Institutional Research (28th, Phoenix, AZ, May 15-18, 1988).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Behavior Rating Scales; *College Students; Higher Education; Institutional Research; *Rating Scales; *Student Attitudes; *Student Evaluation of Teacher Performance; *Test Format; *Test Results; Training
IDENTIFIERS *AIR Forum

ABSTRACT

The effect of rating scale format (behaviorally anchored rating scales (BARS) and Likert) and rater training on leniency and halo in student ratings of instruction was investigated. The subjects (N=269) were students enrolled in required courses at a graduate theological seminary in the southwestern United States. A repeated measures design controlling for teacher and course was used. Findings indicated that: (1) training was effective in reducing leniency and halo in ratings from both instruments; (2) trained raters exhibited less leniency on two rating dimensions when using BARS than when using the Likert scale; and (3) trained raters exhibited less halo when using the Likert than when using the BARS. The findings demonstrate the importance of focusing efforts to improve quality of ratings on the students rather than on the format of the instrument. Contains 24 references. (Author/KM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED298863

Improving the Quality of Student Ratings of Instruction: A Look at Two Strategies

Stuart S. Cook

Director of Institutional Research

Dallas Theological Seminary

3909 Swiss Avenue

Dallas, TX 75204

(214) 824-3094

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

AIR

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

HE 021 854

Running Head: STUDENT RATINGS OF INSTRUCTION

for Management Research, Policy Analysis, and Planning

This paper was presented at the Twenty-Eighth Annual Forum of the Association for Institutional Research held at the Hyatt Regency Phoenix Hotel, Phoenix, Arizona, May 15-18, 1988. This paper was reviewed by the AIR Forum Publications Committee and was judged to be of high quality and of interest to others concerned with the research of higher education. It has therefore been selected to be included in the ERIC Collection of Forum Papers.

Teresa Karolewski
Chair and Editor
Forum Publications Editorial
Advisory Committee

Abstract

The effect of rating scale format (behaviorally anchored rating scales [BARS] and Likert) and rater training on leniency and halo in student ratings of instruction was investigated. The subjects (N = 269) were students enrolled in required courses at a graduate theological seminary in the Southwest United States. A repeated measures design controlling for teacher and course was used. Findings indicated that (a) training was effective in reducing leniency and halo in ratings from both instruments, (b) trained raters exhibited less leniency on two rating dimensions when using BARS than when using the Likert scale, and (c) trained raters exhibited less halo when using the Likert than when using the BARS. The findings demonstrate the importance of focusing efforts to improve quality of ratings on the students rather than on the format of the instrument.

Improving the Quality of Student Ratings of Instruction: A Look at Two Strategies

Introduction

Student ratings have provided the bulk of data used in evaluating college teaching since the first published form for collecting student ratings, the Purdue Rating Scale of Instruction, was released in 1926 (Darr, 1977). In the years following 1926, the practice of collecting student ratings and the variety of methods used to gather them increased substantially. Today administrators use student ratings, collected by a variety of methods, for faculty evaluation on college campuses across the nation.

In a milieu where accountability is a watchword, student ratings have taken on an ever more important role in accounting for the effectiveness of teaching (McKeachie, 1979, p. 394). Miller noted the centrality of a teacher's classroom performance to promotion and tenure decisions and the great influence student evaluation has on these decisions (1987, p. 31). In this climate it is important to be sure that the very best methods are being used to gather and interpret student rating data.

Two imperfections in student ratings of instruction having received extensive attention in the literature are leniency and halo (Jacobs, Kafry, & Zedeck, 1980; Kingstrom & Bass, 1981; Schwab, Heneman, & DeCotiis, 1975). Leniency refers to the tendency of students to mark the rating scale toward the high

end. Halo refers to a tendency to allow overall impressions to guide completion of the evaluation instrument rather than scoring each item independently. The extent to which leniency and halo occur in student ratings determines in part the reliability of the ratings which in turn limits their validity.

The literature documents two strategies to reduce leniency and halo in student ratings: (a) using alternative scale formats (Bernardin, 1977; Bernardin, Alvarez, & Cranny, 1976; Dickinson & Zellinger, 1980; Jacobs, Kafry, & Zedeck, 1980; Keaveny & McGann, 1975; Kingstrom & Bass, 1981; Kinicki, Bannister, Hom, & Denisi, 1985; Reardon & Waters, 1979; Schwab, Heneman, & DeCotiis, 1975) and (b) training students to provide more accurate ratings (Bernardin, 1978; Bernardin & Pence, 1980; Bernardin & Walter, 1977; Borman, 1975, 1979; Ivancevich, 1979; Keeley & Browne, 1978).

Alternative Rating Formats

One of the alternative rating formats intended to reduce leniency and halo error was introduced in 1963 by Smith and Kendall. Scales produced according to the Smith and Kendall procedure are called behaviorally anchored rating scales (BARS) or behavioral expectation scales (BES). For simplicity, the abbreviation BARS will be used throughout the paper to refer to these scales.

BARS are developed through a procedure in which the raters themselves report specific behavioral examples of ratee

performance. These behavioral examples or anchors are collected into separate scales according to performance dimensions and arranged from examples of excellent performance to poor performance. The rater records examples of typical ratee performance and marks the scale where he feels these examples would fall in relation to the scaled anchors. The mean value of the newly scaled examples is used as the score on each rating dimension.

A review of the literature by Schwab, Heneman, and DeCotiis (1975) showed that ratings from BARS studied between 1963 and 1975 had not been proved to be superior to ratings from other scale formats in suppressing psychometric error. Kingstrom and Bass (1981) reviewed twenty-one studies, published between 1965 and 1980, comparing the psychometric properties of BARS and alternative rating formats. Kingstrom and Bass found that the evidence for the superiority in reducing leniency and halo of BARS over other rating formats was inconclusive.

Kinicki, Bannister, Hom, and Denisi (1985) compared a ten-dimension BARS developed by undergraduate university students with the Purdue University Scales, developed by the Measurement and Research Center at Purdue University in 1968. They found that the BARS was superior in control of halo error but inferior in control of leniency error.

In summary, the literature on comparative studies of psychometric error in BARS and other rating formats leaves an

unclear picture. Some studies found BARS superior on halo and/or leniency, some found alternative rating formats superior, and some found no significant differences.

Rater Training

In 1983 Follman, reviewing the state of the art of student rating of instruction, stated that validity of student ratings was a basic issue remaining to be resolved and that rater training might be an answer. He said, "Rater training may be efficacious in reducing leniency and halo effect and thus increase reliability and validity. Substantial additional research is required" (Follman, 1983, p. 9).

Borman (1975) studied the effects of a short training session (5 to 6 minutes) for raters on later scale use. He found that halo error, as measured by the variance across dimensions was reduced after the training.

Bernardin and Walter (1977) conducted an extensive series of training programs and evaluated their impact on later student course ratings. They found that training could have a significant effect in reducing halo and leniency in a BARS instrument.

Bernardin (1978) investigated effects over time of a comprehensive (one-hour) versus an abbreviated (five-minute) rater training session on leniency and halo error in student ratings of instruction. He found that training in reducing leniency and halo given in the second week of the semester was

effective in reducing those errors in ratings taken at the end of the semester. However, the halo and leniency reappeared in ratings taken one academic year later.

Ivancevich (1979) studied managers' ratings made twelve months after training. Ivancevich, like Bernardin, found that the effects of training on leniency and halo errors deteriorated over time.

Borman (1979) investigated the effects of training on student ratings of videotaped performance of managers and recruiters using five different rating scale formats. Borman found that training had a small but significant effect in reducing halo error in a BARS instrument.

Bernardin and Pence (1980) found that training of raters reduced leniency and halo error on BARS but also reduced accuracy. Their study compared the ratings of those trained with the "true" ratings of performance determined by independent, expert raters. Accuracy was operationalized as agreement with the rating of the expert raters. They theorized that rater training that included showing skewed distributions of ratings as evidence of leniency and giving exhortations to conform to more of a normal distribution simply created a new response set among the raters. The raters, they hypothesized, would have a lower response set following training but also would have less accuracy in their ratings. Thus they argued that rater training should focus on accuracy and fairness of ratings.

In summary, the literature demonstrated that rater training could be used to reduce psychometric errors in performance ratings. In particular, training given to students to avoid leniency and halo errors through accurate observation was effective in reducing the presence of these errors in ratings.

Purpose of the Study

BARS never have fulfilled the promise of providing more error-free ratings of instruction, but training of students in doing ratings has been shown effective for this purpose. The BARS format, if used correctly, should elicit more accurate evaluation than a summative scale. The failure of researchers to demonstrate the ability of BARS to reduce psychometric error may be due to the novelty of the scale format (compared to the familiar Likert scale). When students at the end of the semester are given a rating scale that requires careful observation and recording of specific examples of their instructors' performance, it is likely that they will not use the form as intended.

In light of these considerations a study was conducted to validate the conclusion that training improves the psychometric quality of student ratings (regardless of rating format) and to test the effectiveness of the BARS rating format under conditions that favored its use more than had been tested previously. In addition, a new setting for the study, a graduate theological seminary, would provide a basis for generalizing the results of prior research to this education environment.

Method

Hypotheses

The literature clearly indicated that training of raters was effective in reducing leniency and halo. This effect had never been tested in a graduate theological seminary, therefore the first hypothesis was: Ratings following rater training will show less leniency and halo error than ratings without rater training.

The research on BARS as an alternate format to the typical Likert scale indicated that the complexity and unfamiliarity of the form and the need to record observations of teacher performance over a period of time may have biased previous comparisons of the two formats. Therefore the BARS and Likert scales were compared only for the groups who received training. Thus the second hypothesis was: With training BARS will contain less leniency and halo error than Likert scales.

Population

The setting for the study was a graduate theological seminary in the Southwest United States. The institution had a policy of collecting student ratings of instruction from all students in all courses. The student population was primarily male (89.2%) and married (64.2%). A total of 269 students participated in the study.

Experimental Design

Since ratings of instruction vary by teacher and course, the experimental design had to control for these two confounding variables. The problem was solved by choosing for the study students in four required courses that had two sections taught by the same professor. The four courses were selected at random from all such courses being taught. Analysis of the previous semester's ratings of instruction indicated no significant differences in the ratings of different sections of required courses taught by the same professor.

The design of the study followed Kirk's Split Plot Factorial 2 by 4 by 2 design (SPF 24.2) (1982, p. 523). The two between blocks treatments were rater training (training and no training) and course (four courses). The within blocks treatment was type of rating scale (BARS and Likert). A repeated measures design was used in which the individual student composed a block.

Treatments

During the tenth week of a regular fifteen-week semester the subjects chosen for the study received the treatments described below. Both treatments took place during regular classtime with the permission of the course instructor.

Training sections. The students in sections receiving training listened to a fifteen-minute presentation on halo and leniency error and were given a copy of the rating instruments to be used in rating the professor in that course. The presentation

stressed the importance and benefit of student ratings and the need to avoid committing leniency or halo errors. The students were given actual copies of the BARS and Likert rating instruments (with a cover sheet for note-taking during the presentation) to keep with their classnotes for the rest of the semester. They were asked to begin to record observations about the instructor's behavior on the six BARS scales. They were told to retain the rating instruments as they would be collected at the end of the semester.

No training sections. The students in sections not receiving the training were presented with the currently-used Likert instrument and asked to critique it. They were told that the administration desired to improve the instrument and that their comments were sought. The students were given the opportunity to ask the researcher questions about the student rating process or about the instrument itself. No instruction was given regarding errors in the student rating process. All instruments were collected from the students at the end of fifteen minutes. The students were told that the researcher would be back in their class at the end of the semester to collect their ratings on that course.

Instrumentation

The Likert scale was a fourteen-item instrument developed and used at the institution where the study took place. This

instrument was used without modification as the Likert instrument.

Since the purpose of the study was to compare a Likert and a BARS instrument on separate dimensions of instruction, the ratings of the previous semester from Likert instrument were analyzed by means of principal components analysis to determine the dimensions measured. Analysis indicated that the Likert instrument measured six readily identifiable dimensions of instruction. In conjunction with the nature of the items that loaded on each factor and the research literature on the dimensions of instruction, the six factors were labeled Creating Interest (CI), Grading and Feedback (GF), Establishing Rapport (ER), Course Preparation (CP), Subject Coverage (SC), and Workload (WL). From this analysis the equations were developed for transforming the 14 raw scores into 6 factor scores.

To develop a BARS instrument that contained dimensions parallel to the Likert scale, the recommended procedures of Bernardin, LaShells, Smith, and Alvares (1976) were followed. They recommended four steps as follows. First, a group of students develop a pool of statements of teacher behavior that represent excellent, average, and poor performance on each of the dimensions of teaching to be studied. Second, a different group of students retranslates each behavior statement into the dimension where they believe the statement fits. Statements that are placed in the same category by 60% of the group are retained.

Third, another group rates the surviving statements on their desirability on a scale from one to seven. Statements with a standard deviation of scale value of 2.0 or less are retained. Fourth, the scales are constructed as a vertical axis with behavioral anchors placed along the axis, marked on the axis to indicate their scale value. Table 1 displays the statistics representing the results of these development procedures.

Insert Table 1 about here.

Analysis

Preparation of data. Scores on the 14 Likert items and the 6 BARS were collected for 269 students (124 students in sections receiving training and 145 in sections not receiving training). Using the factor equations mentioned above the 14 Likert items were transformed into 6 scale scores corresponding to the 6 BARS scores. The BARS value range was 10 to 70 while the Likert scale scores were standard scores with mean 0 and standard deviation 1. To allow comparison of the scores from the two scale types, the Likert scale scores were transformed scale by scale to have the same mean and standard deviation as the BARS. This procedure precluded comparisons of the grand mean of Likert and BARS (they were equal by definition), but it allowed testing of the hypotheses of interest to the study. The scores themselves were used to test leniency.

Measures of halo for Likert scores and BARS scores were computed as the standard deviation across the six scale scores for each scale type. Halo score is inverse to halo error, that is, a higher score on halo indicates less halo error.

Hypothesis 1. Table 2 displays the means and standard deviations for leniency on the six rating dimensions on both instruments and halo across all six dimensions on both instruments. Leniency was lower on all six scales for the students who received training than for the students who did not. Multivariate test for the effect of training on leniency error in the six scales yielded a value of Hotellings T^2 = .1721, which is equal to F (6, 256) = 7.3427, $p < .001$. Thus it was found that training reduced leniency on the six rating dimensions taken together. Univariate F-tests with df (1, 261) were performed on the six scales separately. All six tests yielded results with $p < .09$, four scales (CI, CP, GF, and WL) yielded results with $p < .05$.

Insert Table 2 about here.

Table 2 shows that there was less halo error (larger halo value) in the ratings provided by the trained raters than by the untrained raters. The repeated measures ANOVA with halo as dependent measure yielded a value of F (1, 261) = 13.44,

$p < .001$ for the training effect. Thus the training session was considered effective in reducing halo error.

Hypothesis 2. Table 3 displays the means and standard deviations for leniency on the six rating dimensions on both instruments and halo across all six dimensions on both instruments for the training sections only. A multivariate test was performed to determine the significance of the difference between means for the six leniency dimensions. The value of Hotellings T^2 that resulted was .1429, which is equivalent to $F(6,256) = 6.098$, $p < .001$. This result indicated that there was a significant effect for scale among the students receiving training on the multivariate. The data of Table 3 indicate that the dimensions measured with the BARS contained less leniency than those measured with the Likert. Univariate follow-up tests on the six dimensions showed that only on two scales (SC and WL) was the difference significant ($p < .05$).

Insert Table 3 about here.

The test for halo error yielded a significant effect for scale type, $F(1, 261) = 37.03$, $p < .001$, and no interaction between training and scale type, $F(1, 261) = .42$, $p = .518$, indicating difference in halo for scale type regardless of presence or absence of training. Table 3 indicates that this

significant difference favored the Likert scales (less halo error indicated by larger halo value) over the BARS.

Discussion

The results of this study confirmed what has been reported in the research literature concerning student ratings of instruction. The main findings were (a) training students improves the psychometric quality of the rating data and (b) the BARS format does not provide consistently more error-free data than the simpler Likert scale.

The literature provides examples of a variety of training strategies for improving student ratings. They all have been shown to be effective in reducing psychometric error. While one might tend to suspect that the Hawthorne effect is causing the improvement, the studies reviewed and the current study were well-controlled against such threats to external validity. When students are reminded of the importance of ratings and given information and opportunity to provide high-quality rating data, they do so with impressive regularity.

This researcher is ready to nail down the coffin lid on BARS as a viable format for student ratings. This study confirmed the inability of this rating format to elicit more error-free data than the Likert scale. In addition, the response of the students to the BARS format in the current study was less than enthusiastic. Though some students saw the value of the more complex and behavior-specific BARS, most felt it was too much

work to expect students to complete every semester for every course.

The four faculty members who participated in the study, likewise, felt the feedback from the Likert scale was just as informative as were the BARS results. They uniformly expressed their preference for the written comments requested on the standard institutional (Likert) form. In their opinion the Likert responses with written comments were just as valuable as the specific observations recorded on the BARS.

References

Bernardin, H. J. (1977). Behavioral expectation scales versus summated scales: A fairer comparison. Journal of Applied Psychology, 62, 422-427.

Bernardin, H. J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. Journal of Applied Psychology, 63, 301-308.

Bernardin, H. J., Alvares, K. M., & Cranny, C. J. (1976). Recomparison of behavioral expectation scales to summated scales. Journal of Applied Psychology, 61, 564-570.

Bernardin, H. J., LaShells, M. B., Smith P. C., & Alvares, K. M. (1976). Behavioral expectation scales: Effects of developmental procedures and formats. Journal of Applied Psychology, 61, 75-79.

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. Journal of Applied Psychology, 65, 60-66.

Bernardin, H. J., & Smith, P. C. (1981). A clarification of some issues regarding the development and use of behaviorally anchored rating scales (BARS). Journal of Applied Psychology, 66, 458-463.

Bernardin, H. J., & Walter, C. S. (1977). Effects of rater training and diary-keeping on psychometric error in rating. Journal of Applied Psychology, 62, 64-69.

Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 60, 556-560.

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. Journal of Applied Psychology, 64, 410-421.

Darr, R. F., Jr. (1977). Evaluation of college teaching: State of the art, 1977. Paper presented at the Ohio Academy of Science Psychology Division, Columbus, OH. (ERIC Document Reproduction Service No. ED 162 559)

Dickinson, T. L. & Zellinger, P. M. (1980). Comparison of the behaviorally anchored rating and mixed standard scale formats. Journal of Applied Psychology, 65, 147-54.

Follman, J. (1983, March). Student ratings of faculty teaching effectiveness: Revisited. Paper presented at the annual meeting of the Association for the Study of Higher Education, Washington, DC. (ERIC Document Reproduction Service No. ED 232 556)

Ivancevich, J. M. (1979). Longitudinal study of the effects of rater training on psychometric errors in ratings. Journal of Applied Psychology, 64, 502-508.

Jacobs, R., Kafry, D., & Zedeck, S. (1980). Expectations of behaviorally anchored rating scales. Personnel Psychology, 33, 595-640.

Keaveny, T. J., & McGann, A. F. (1975). Comparison of behavioral expectation scales and graphic rating scales. Journal of Applied Psychology, 60, 695-703.

Keeley, S. M., & Browne, M. N. (1978). Improving student evaluation forms: Related need for critical thinking emphasis and trained raters, Peabody Journal of Education, 55, 305-308.

Kingstrom, P. O., & Bass, A. R. (1981). A critical analysis of studies comparing behaviorally anchored rating scales (BARS) and other rating formats. Personnel Psychology, 34, 263-289.

Kinicki, A. J., Bannister, B. D., Hom, P. W., & Denisi, A. S. (1985). Behaviorally anchored rating scales vs. summated rating scales: Psychometric properties and susceptibility to rating bias. Educational and Psychological Measurement, 45, 535-549.

Kirk, R. E. (1982). Experimental design: Procedures for the behavioral sciences (2nd ed.). Belmont, CA: Brooks/Cole.

McKeachie, W. J. (1979). Student ratings of faculty: A reprise. Academe, 65, 384-397.

Miller, R. I. (1987). Evaluating faculty for tenure and promotion. San Francisco: Jossey-Bass.

Reardon, M., & Waters, L. K. (1979). Leniency and halo in student ratings of college instructors: A comparison of three rating procedures and implications for scale validity. Educational and Psychological Measurement, 39, 159-162.

Schwab, D. P., Heneman, H. G., III, & DeCotiis, T. A. (1975). Behaviorally anchored rating scales: A review of the literature. Personnel Psychology, 28, 549-5

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.

Table 1

Descriptive Statistics for BARS Dimensions

	Development steps					
	1	2	3		4	
Dimension	N	N	SDs	N	Means	SDs
CI	81	32	.29-2.06	6	1.43-6.50	.65-1.41
CP	59	40	.52-1.97	7	1.50-6.38	.67-1.81
ER	68	50	.52-1.73	7	1.44-6.50	.52-1.29
GF	70	60	.29-1.68	7	1.33-6.25	.49-1.45
SC	42	20	.62-1.56	5	2.17-6.12	.62-1.56
WL	40	22	.68-1.56	7	1.67-6.17	.72-1.21

Table 2

Means and Standard Deviations for Hypothesis 1

	Training (N = 124)		No Training (N = 145)	
Dimension	Mean	SD	Mean	SD
CI	52.86	9.42	56.95	8.47
CP	46.41	11.73	49.91	11.00
ER	54.62	8.22	57.09	7.23
GF	40.51	12.48	42.95	12.79
SC	51.61	11.14	54.23	9.38
WL	51.65	11.06	55.78	9.53
Halo	12.18	4.12	11.11	4.13

Table 3

Means and Standard Deviations for Hypothesis 2 (Training Sections Only, N = 124)

	Likert		BARS	
Dimension	Mean	SD	Mean	SD
CI	53.55	12.67	52.16	11.10
CP	46.54	14.61	46.28	13.66
ER	54.78	10.85	54.45	10.10
GF	41.40	15.11	39.61	13.73
SC	53.11	13.50	50.11	12.12
WL	53.24	13.58	50.06	12.60
Halo	12.76	5.83	10.01	3.76